Short-term practice effects in mild cognitive impairment: Evaluating different methods of change

Kevin Duff[a,b], **Taylor J. Atkinson**[a], **Kayla R. Suhrie**[a], **Bonnie C. Allred Dalley**[a], **Sydney Y. Schaefer**[b,c], and **Dustin B. Hammers**[a,b]

[a]Department of Neurology, Center for Alzheimer's Care, Imaging and Research, University of Utah, Salt Lake City, UT, USA

[b]Center on Aging, University of Utah, Salt Lake City, UT, USA

[c]School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ, USA

Abstract

Practice effects are improvements on cognitive tests as a result of repeated exposure to testing material. However, variability exists in the literature about whether patients with amnestic mild cognitive impairment (MCI) display practice effects, which may be partially due to the methods used to calculate these changes on repeated tests. The purpose of the current study was to examine multiple methods of assessing short-term practice effects in 58 older adults with MCI. The cognitive battery, which included tests of memory (Hopkins Verbal Learning Test–Revised and Brief Visuospatial Memory Test–Revised) and processing speed (Symbol Digit Modalities Test and Trail Making Test Parts A and B), was administered twice across one week. Dependent t tests showed statistically significant improvement on memory scores ($ps < .01$, $ds = 0.8$–1.3), but not on processing speed scores. Despite this, the sample showed no clinically meaningful improvement on any cognitive scores using three different reliable change indices. Regression-based change scores did identify relatively large groups of participants who showed smaller than expected practice effects, which may indicate that this method is more sensitive in identifying individuals who may portend a declining trajectory. Practice effects remain a complex construct, worthy of continued investigation in diverse clinical conditions.

Keywords

Cognitive change; Memory; Mild cognitive impairment; Practice effects; Reliable change index

Practice effects are improvements on cognitive tests due to repeated exposure to the testing materials. Although these improvements in test scores may not reflect true change in cognitive abilities, and they have traditionally been dismissed as an artifact of the testing situation, it remains possible that practice effects do inform us about unique aspects of cognition (Duff, Callister, Dennett, & Tometich, 2012). For example, smaller than expected

practice effects in older adults may herald a declining trajectory (Duff et al., 2011), poorer response to an intervention (Duff, Beglinger, Moser, Schultz, & Paulsen, 2010), or greater risk of Alzheimer's-related pathology (Duff, Foster, & Hoffman, 2014; Galvin et al., 2005; Mormino et al., 2014). The potential for practice effects to inform clinicians and researchers about cognitive course, treatment response, and brain pathology has also been examined in neurodegenerative disorders (Duff et al., 2007), traumatic brain injury (Rogers, Fox, & Donnelly, 2015), and stroke (Chiu et al., 2014). Ultimately, practice effects may provide valuable information in clinical (e.g., accurately identifying early cognitive decline) and research (e.g., identifying at-risk subjects for clinical trials) settings.

Practice effects tend to be quite robust in healthy older adults (Calamia, Markon, & Tranel, 2012; McCaffrey, Duff, & Westervelt, 2000) and generally absent in patients with dementia (Cooper et al., 2001; Helkala et al., 2002; Zehnder, Bläsi, Berres, Spiegel, & Monsch, 2007). Perhaps not surprisingly, findings of practice effects in patients with mild cognitive impairment (MCI) have been equivocal. For example, some studies have reported an absence of practice effects in MCI on various cognitive measures across various retest intervals (Britt et al., 2011; Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004; Darby, Maruff, Collie, & McStephen, 2002; Schrijnemaekers, de Jager, Hogervorst, & Budge, 2006). Conversely, others have reported improvements on repeated testing in these patients (Duff et al., 2007; Mathews et al., 2014; Yan & Dick, 2006), and that patients with MCI who do not show practice effects tend to have worse outcomes than those that do show improvements on retesting (Duff et al., 2011; Hassenstab et al., 2015; Machulda et al., 2013).

Although there may be many reasons for these equivocal findings of practice effects in MCI (e.g., different tests used, different retest intervals, different levels of cognitive impairment in samples, different ages and education levels in samples), the discrepant findings may also be due to various methods for quantifying, calculating, or examining practice effects in these cohorts. Most of the above referenced studies examine simple differences in raw cognitive test scores to determine whether practice effects occur for their entire sample, often followed by dependent/paired t tests or repeated measures analysis of variance (ANOVA) to statistically evaluate the effect. These methods best determine whether practice effects occur at a group level. Conversely, in clinical neuropsychology, more sophisticated mathematical formulae are used to determine whether the cognitive change in an individual patient is significantly different than the change seen in some normative sample (Duff, 2012; Hinton-Bayre, 2010; Maassen, Bossema, & Brand, 2009). These methods best determine how much improvement occurs relative to expectations. In this approach, two mathematical formulae seem to dominate: reliable change index (RCI) and standardized regression-based (SRB) models. In the classic RCI that corrects for practice effects (Chelune, Naugle, Lüders, Sedlak, & Awad, 1993), the difference between two test scores (Times 1 and 2) for an individual patient is compared to the difference between these two test scores for some normative group. The resulting z score indicates how much change this individual made compared to the normative group. Revisions to the RCI have been made, including Iverson's RCI (Iverson, 2001), which seems to be most in favor and is presented in Table 1. In the classic SRB (McSweeny, Naugle, Chelune, & Lüders, 1993), multiple regression is used to predict a Time 2 score using the Time 1 score, and this is referred to as the "simple SRB." This model can be expanded by including other possibly relevant clinical information (e.g.,

age, education, retest interval), and this is referred to as the "complex SRB." In both simple and complex SRBs, the predicted Time 2 score is compared to the observed Time 2 score to yield a z score, which indicates how large a deviation of change from normal has occurred in this individual patient (see Table 1).

For both the RCIs and the SRBs, z scores between −1.645 and +1.645 indicate "no change" or "stability." Z scores below −1.645 typically indicate significant "decline," and z scores above +1.645 indicate significant "improvement." This demarcation point of ±1.645 was originally chosen because of its parallel with traditional parametric statistical testing, but there is little in the way of data to support it as the best cut-point for assessing change (Duff, 2012). Improvements of +1.53 or declines of −1.18 would appear to still tell us something about change, even though they technically fall within the "no change" range. In general, negative z scores seem to indicate smaller than expected change relative to the comparison group, whereas positive z scores indicate larger than expected change. Admittedly, this is a complex concept, as a negative z score may represent one of two possibilities. First, if extreme enough, a negative z score may indicate an actual decline in scores (e.g., Delayed Recall score on the Hopkins Verbal Learning Test–Revised dropping from 8 at baseline to 5 at follow-up). Alternatively, a negative z score may not reflect actual decline, but a diminished practice effect relative to the comparison sample (e.g., Delayed Recall score on the Hopkins Verbal Learning Test–Revised improving from 8 at baseline to 9 at follow-up— improvement of 1 point—when the comparison group improves by 2.2 points across the same time period). Across longer retest intervals (e.g., weeks, months, years), negative z scores are more likely to reflect actual decline in scores. Across shorter retest intervals (e.g., hours, days), negative z scores are more likely to reflect diminished improvement relative to the comparison sample. Positive z scores are somewhat easier to grasp, as they always reflect more improvement relative to the comparison group.

When comparing the different mathematical formulae, there is no clear consensus as to which method (e.g., RCI correcting for practice effects, simple SRB, complex SRB) is preferred to detect reliable and clinically meaningful change. For example, even though Temkin, Heaton, Grant, and Dikmen (1999) reported that their actual data did not find large differences between multiple reliable change indices, they recommended more complex SRBs. In applying Temkin's results to four clinical samples, Heaton et al. (2001) found no superiority of complex SRB to RCI that corrects for practice. Similarly, Frerichs and Tuokko (2005) found little difference between methods in their actual data, but recommended that RCI that corrects for practice and aging might be worth pursuing. Using simulated and actual data, Maassen et al. (2009) advocated that some reliable change formulae are preferred to others depending on the seriousness of making Type 1 or Type 2 errors. More recently, Hinton-Bayre (2010) examined multiple reliable change methods with actual data and concluded "there was apparently no universally more sensitive or conservative [reliable change] model" (p. 252). Anecdotally, it seems reasonable to assume that the complex SRB, which incorporates baseline cognitive functioning, demographic information, and other variables, might yield more sensitive information about change. Nonetheless, empirical data are lacking. To our knowledge, no studies have examined these various methods of identifying cognitive change in patients with MCI, especially across brief retest intervals. Therefore, the purpose of this study was to compare multiple methods of examining change

across one week in patients with MCI to evaluate whether practice effects occur in this group. Although this was largely a descriptive study, it was hypothesized that the complex SRB method would better identify change in these patients than the dependent *t* tests, RCI correcting for practice, or simple SRB. A secondary purpose of these analyses was to examine the RCI and SRB change norms developed by Duff (2014) in an independent sample of patients with MCI.

Method

Participants

Fifty-eight older adults were recruited from either a cognitive disorders clinic or presentations at independent living facilities and senior centers for a research project on cognitive training. They had a mean age of 75.8 (*SD* = 5.8) years, and ranged from 65 to 89 years old. They were evenly divided by sex. Mean education was 16.5 years (*SD* = 2.7), and premorbid intellect was average (Wide Range Achievement Test–4, WRAT–4, Reading: *M* = 109.3, *SD* = 8.0) (Wilkinson & Robertson, 2006). For inclusion in the study, all participants met criteria for mild cognitive impairment–amnestic subtype (i.e., subjective memory complaints, objective memory problems, no significant functional limitations). Subjective memory complaints were reported by participants and/or knowledgeable collaterals. Objective memory problems were determined by a significant discrepancy between delayed recall measures and an estimate of pre-morbid intellect (see below). Minimal to no functional limitations (e.g., still driving, managing medications, handling finances, completing household chores) were reported by participants and/or knowledgeable collaterals. Additionally, participants had to be 65 years or older and have adequate vision, hearing, and motor abilities to complete cognitive testing. Exclusion criteria included medical comorbidities likely to affect cognition (e.g., history of major neurological disorders, major psychiatric disorders, or substance abuse); use of anticonvulsant or antipsychotic medications; severe depression as indicated by scores of >6 on 15-item Geriatric Depression Scale (GDS) or >14 on 30-item GDS; and residing in a skilled-nursing facility.

Procedure

All procedures were approved by the local Institutional Review Board before the study commenced. All participants provided informed consent before completing any procedures. The following measures were administered at a baseline visit:

1. Hopkins Verbal Learning Test–Revised (HVLT–R; Brandt & Benedict, 2001) is a verbal learning task of 12 words over three learning trials, with correct words summed for the Total Recall score (range = 0–36). The Delayed Recall score is the number of correct words recalled after a 20–25 min delay (range = 0–12). For all HVLT–R scores, higher values indicate better performance.

2. Brief Visuospatial Memory Test–Revised (BVMT–R; Benedict, 1997) is a visual learning task of six geometric designs in six locations on a card, with correct designs and locations summed for the Total Recall score (range = 0–36). The Delayed Recall score is the number of correct designs and locations recalled

after a 20–25-min delay (range = 0–12). For all BVMT–R scores, higher values indicate better performance.

3. Symbol Digit Modalities Test (SDMT; Smith, 1973) is a divided attention and psychomotor speed task, with the number of correct responses in 90 seconds being the total score (range = 0–110), and higher values indicating better performance.

4. Trail Making Test Parts A and B (TMT–A, TMT–B; Reitan, 1992) are tests of visual scanning/processing speed and set shifting, respectively. For each part, the score is the time to complete the task, and higher values indicate poorer performance.

5. WRAT–4 Reading (Wilkinson & Robertson, 2006) is used as an estimate of premorbid intellect, in which an individual attempts to pronounce irregular words. The score is standardized ($M = 100$, $SD = 15$) to age-matched peers, with higher values indicating better performance.

6. Geriatric Depression Scale (GDS; Yesavage et al., 1983) is a 30-item screening measure of depressive symptoms in the elderly. Higher scores indicate more depressive symptoms.

After approximately one week ($M = 7.1$ days, $SD = 0.6$, range = 6–9), the HVLT–R, BVMT–R, SDMT, TMT–A, and TMT–B were repeated. The same form of each test was used to maximize practice effects.

Analyses

Four methods were used to examine practice effects across one week in this sample. First, dependent t tests were calculated for each of the repeated cognitive measures, using the raw scores at baseline and one-week visits. Cohen's d was calculated as $t/\sqrt{n}$ (Rosenthal, 1991). Second, three different change formulae were calculated for each participant: Iverson's RCI (Iverson, 2001), which corrects for practice effects; simple SRB, which uses the baseline score to predict the one-week score; and complex SRB, which uses the baseline score and demographic variables (e.g., age, education, gender) to predict the one-week score. The RCI and two SRB scores were calculated using the data presented in Duff (2014), which examined these same cognitive tests across a one-week retest interval in 167 nondemented older adults. It should be noted that the majority of this comparison sample were cognitively intact (56%), but that a minority were classified as having amnestic MCI (44%). Demographically, the two samples were very similar (e.g., mean age: Duff, 2014 = 78.6 years, current = 75.8 years; mean education: Duff, 2014 = 15.4 years, current = 16.5 years; all Caucasian in both samples). They were also comparable on an estimate of pre-morbid intellect (mean WRAT–4 Reading: Duff, 2014 = 107.8, current = 109.3). The Duff (2014) sample did have more females than the current sample (81% vs. 50%). The resulting RCI and two SRB z scores were compared in a repeated measures ANOVA, with Fisher's least significant difference test as the post hoc comparison. Third, the RCI and two SRB z scores were trichotomized ("decline," or smaller than expected levels of improvement relative to the comparison sample ≤ −1.645; "stable," or comparable levels of improvement relative to the comparison sample = −1.644 to +1.644; "improve," or larger than expected levels of

improvement relative to the comparison sample +1.645), which is typical in this methodology (Duff, 2012). If the z scores were normally distributed, then one would expect that 5% of participants would show "decline," 90% would remain "stable," and 5% would "improve." A one-sample chi-square test examined whether a change formula identified 90% of cases as remaining stable. Chi-square tests were repeated for each change formula for each cognitive test score. Fourth, the RCI and two SRB z scores were dichotomized as "stable" (i.e., $z = -1.644$ to $+1.644$) or "not stable" (z −1.645 or z +1.645). These dichotomized change scores were compared across the three change formulae (RCI, simple SRB, complex SRB) using Cochran's Q test, which is a nonparametric test to examine three or more matched sets of frequencies or proportions (Linebach, Tesch, & Kovacsiss, 2014). This test allows us to examine whether there were differences in the three change formulae in their ability to identify stable/not stable cases. If overall differences were identified with the Cochran Q test, then post hoc McNemar's tests were used to see which of the formulae were significantly different from the others.

Results

Examining practice effects with dependent t tests

Dependent t tests showed statistically significant improvement in raw scores on all four memory scores: HVLT–R Total Recall, $t(57) = -6.20$, $p < .001$, $d = 0.81$, HVLT–R Delayed Recall, $t(57) = -7.65$, $p < .001$, $d = 1.01$, BVMT–R Total Recall, $t(57) = -9.62$, $p < .001$, $d = 1.26$, and BVMT–R Delayed Recall, $t(57) = -9.26$, $p < .001$, $d = 1.22$. Conversely, there were no statistically significant differences on any of the processing speed scores: SDMT, $t(57) = -0.09$, $p = .93$, TMT–A, $t(57) = 0.34$, $p = .74$, or TMT–B, $t(55) = 1.66$, $p = .10$. See Table 2.

Examining practice effects with change formulae and repeated measures ANOVA

The mean Iverson RCI, simple SRB, and complex SRB scores for the entire sample are presented in Table 3. These values are z scores, with $M = 0$, $SD = 1$. On average, the negative z scores indicate smaller than expected improvement across one week compared to a normative group (z scores for TMT–A and TMT–B were reversed, so negative z scores also indicate smaller than expected improvement). This may be best exemplified by the Iverson RCI, which corrects for practice effects in the comparison sample. In Duff (2014), the comparison sample improved by 2.2 points on the Delayed Recall of the HVLT–R across one week. In the current sample, they averaged 2.1 points improvements on this same cognitive variable across this same retest interval. So, even though the current sample improves on the HVLT–R Delayed Recall trial, their improvement was smaller than expected, as indicated by the mean RCI z score of −0.05. The simple SRB, which also corrects for baseline cognitive level, shows that the same amount of improvement is even smaller than expected (i.e., mean z score of −0.89). The complex SRB, which also corrects for demographic information, showed an even smaller than expected performance (i.e., mean z score of −1.09).

Repeated measures ANOVAs for each cognitive test score indicated statistically significant differences between the change formulae: HVLT–R Total Recall, $F(1, 57) = 57.5$, $p < .001$,

$\eta_p^2 = .50$; HVLT–R Delayed Recall, $F(2, 56) = 48.7$, $p < .001$, $\eta_p^2 = .64$; BVMT–R Total Recall, $F(2, 56) = 34.0$, $p < .001$, $\eta_p^2 = .55$; BVMT–R Delayed Recall, $F(2, 56) = 13.6$, $p < .001$, $\eta_p^2 = .33$; TMT–A, $F(2, 56) = 17.9$, $p < .001$, $\eta_p^2 = .39$; TMT–B, $F(2, 54) = 13.8$, $p < .001$, $\eta_p^2 = .34$; and SDMT, $F(2, 56) = 41.4$, $p < .001$, $\eta_p^2 = .60$. The post hoc comparisons between change formulae are also presented in Table 3. In general, the complex SRB tended to yield larger z scores than the other two change methods, indicating more change across one week relative to expectations based on a normative group.

Examining practice effects with change formulae and chi-square

Table 4 presents the percentage of cases that decline, remain stable, or improve across one week for the RCI and two SRB scores after z scores were trichotomized (decline   −1.645, stable = −1.644 to +1.644, improve   +1.645). If the z scores were normally distributed, then one would expect that 5% of participants would show "decline" (i.e., smaller than expected improvement), 90% would remain "stable" (i.e., expected levels of improvement), and 5% would show "improvement" (i.e., larger than expected improvement). A cursory look at the percentages reveals different patterns across the change formulae. For example, on the HVLT–R Delayed Recall, the RCI seems to indicate that no participants declined, nearly all remained stable, and fewer than expected improved. Conversely, the simple and complex SRB identified that many more participants declined on the HVLT–R Delayed Recall (34–36%), fewer remained stable, and none improved.

One-sample chi-square tests of the RCI showed no statistically significant differences from the expected number of stable participants (90%) on any of the measures assessed. Specifically, no statistical significance was observed for HVLT–R Total Recall, $\chi^2(1) = 0.01$, $N = 58$, $p = .93$; HVLT–R Delayed Recall, $\chi^2(1) = 2.77$, $N = 58$, $p = .10$; BVMT–R Total Recall, $\chi^2(1) = 0.01$, $N = 58$, $p = .93$; BVMT–R Delayed Recall, $\chi^2(1) = 1.50$, $N = 58$, $p = .22$; TMT–A, $\chi^2(1) = 0.93$, $N = 58$, $p = .34$; TMT B, $\chi^2(1) = 2.29$, $N = 56$, $p = .13$; or SDMT, $\chi^2(1) = 0.28$, $N = 58$, $p = .60$.

One-sample chi-square tests of the simple SRB showed statistically significant differences from the expected number of stable participants (90%) on HVLT–R Total Recall, $\chi^2(1) = 7.36$, $N = 58$, $p = .007$, HVLT–R Delayed Recall, $\chi^2(1) = 38.63$, $N = 58$, $p < .001$, and TMT–B, $\chi^2(1) = 10.87$, $N = 56$, $p = .004$. There were no statistically significant differences from the expected number of stable participants on BVMT–R Total, $\chi^2(1) = 0.12$, $N = 58$, $p = .73$, BVMT–R Delayed Recall, $\chi^2(1) = 0.93$, $N = 58$, $p = .34$, TMT–A, $\chi^2(1) = 1.96$, $N = 58$, $p = .16$, or SDMT, $\chi^2(1) = 0.93$, $N = 58$, $p = .16$.

One-sample chi-square tests of the complex SRB showed statistically significant differences from the expected number of stable participants (90%) on HVLT–R Total Recall, $\chi^2(1) = 7.36$, $N = 58$, $p = .007$, HVLT–R Delayed Recall, $\chi^2(1) = 44.26$, $N = 58$, $p < .001$, BVMT–R Total Recall, $\chi^2(1) = 9.93$, $N = 58$, $p = .002$, BVMT–R Delayed Recall, $\chi^2(1) = 7.36$, $N = 58$, $p = .007$, TMT–B, $\chi^2(1) = 8.13$, $N = 56$, $p = .001$, and SDMT, $\chi^2(1) = 12.88$, $N = 58$, $p < .001$. There was no statistically significant difference from the expected number of stable participants on TMT–A, $\chi^2(1) = 3.38$, $N = 58$, $p = .07$.

Examining practice effects with change formulae and Cochran Q

Table 5 presents the percentage of cases that were classified as stable or not stable across one week for the RCI and two SRB scores after z scores were dichotomized (stable = −1.644 to +1.644, not stable −1.645 or +1.645). Cochran Q tests showed statistically significant differences between the three change formulae for HVLT–R Total Recall, $\chi^2(2) = 9.0$, $N = 58$, $p = .01$, HVLT–R Delayed Recall, $\chi^2(2) = 29.8$, $N = 58$, $p < .001$, BVMT–R Total Recall, $\chi^2(2) = 14.2$, $N = 58$, $p = .001$, BVMT–R Delayed Recall, $\chi^2(2) = 11.1$, $N = 58$, $p = .004$, and SDMT, $\chi^2(2) = 12.3$, $N = 58$, $p = .002$. Post hoc McNemar's tests revealed that: the RCI yielded more stable cases than the simple SRB, which yielded more stable cases than the complex SRB for the HVLT–R Delayed Recall ($p < .001$ for each comparison); the simple SRB yielded more stable cases than the complex SRB for the BVMT–R Total Recall, $p = .008$; the RCI yielded more stable cases than the complex SRB for the BVMT–R Delayed Recall, $p = .004$; and the RCI and the simple SRB yielded more stable cases than the complex SRB for the SDMT, $p = .016$ and $.031$, respectively. Even though the Cochran Q test was statistically significant for the HVLT–R Total Recall, none of the McNemar's tests were ($p > .05$ for each comparison).

Discussion

The purpose of the present study was to compare multiple methods of change in patients with MCI to determine whether short-term practice effects were present in this cohort. When utilizing dependent t tests on raw test scores at baseline and one week, large and statistically significant practice effects were observed on the immediate and delayed recall trials of both memory measures. Conversely, minimal practice effects were seen on the tests of processing speed in this sample when using dependent t tests. When utilizing three different change formulae typically used in clinical neuropsychology (RCI correcting for practice effects, simple SRB, and complex SRB), smaller than expected practice effects were observed for nearly all test scores in this study. Overall, these findings indicate that the method used to examine practice effects (dependent t tests, RCI, SRB) appears to influence whether practice effects are observed and their magnitude.

Dependent t tests are frequently used to compare individuals on the same measure at two time points. Considering just the dependent t test results in the current study, notable practice effects on the memory measures were observed, which would be consistent with others who have observed improvements in test scores with repeated assessments (Duff et al., 2007; Mathews et al., 2014; Yan & Dick, 2006). In some ways, large practice effects on memory tests might be surprising since these patients suffered from primarily memory deficits. But when practice effects have been observed in MCI, they were most robust on memory measures (Duff et al., 2007; Machulda et al., 2013; Yan & Dick, 2006). The absence of practice effects on the processing speed tests would also be supportive of existing studies that have failed to find improvements in patients with MCI (Britt et al., 2011; Cooper et al., 2004; Darby et al., 2002; Schrijnemaekers et al., 2006). These seemingly discrepant findings might be expected given the meta-analysis of Calamia et al. (2012), who noted that practice effects vary depending on multiple factors, including the types of cognitive domains and tests examined. For example, even though improvements were observed on both

learning and delayed recall trials of the HVLT–R and BVMT–R, the effect sizes were largest for the BVMT–R. One possibility for these larger practice effects on the BVMT–R might be its novelty. Recalling the designs and their correct locations on the BVMT–R may be more novel than recalling the list of words on the HVLT–R, and Suchy, Kraybill, and Franchow (2011) have shown that novelty influences practice effects.

Whereas dependent *t* tests are a common method of examining cognitive change in a sample, and they seem to indicate that improvements can occur in patients with MCI, they tell us little about the magnitude of change relative to expectation based on some normative group. Therefore, we also examined practice effects with three change formulae typically used in clinical neuropsychology. The RCI correcting for practice effects (Iverson, 2001) indicated that these memory-impaired patients were showing smaller than expected practice effects across one week compared to a large sample of demographically similar nondemented older adults tested twice with a similar cognitive battery across a similar retest interval (Duff, 2014). With a mean *z* score across the seven measures of –0.21 (after reversing the *z* scores for TMT–A and TMT–B), the practice effects in this cohort was about at the 42nd percentile of those in Duff's sample. The simple SRB (McSweeny et al., 1993), which predicts the Time 2 score from the Time 1 score, also showed smaller than expected practice effects in these individuals with MCI. The magnitude of change was double that found with the RCI correcting for practice effects (–0.45 vs. –0.21), or falling at about the 33rd percentile of Duff's cohort. Finally, the complex SRB, which adds demographic variables to the prediction of a Time 2 score, showed an even smaller than expected practice effect (–0.62 or 26th percentile of Duff's sample). So, although dependent *t* tests show that individuals with MCI do demonstrate practice effects across one week, these sophisticated change formulae indicate that the amount of improvement tends to be smaller than expected relative to their peers.

As a group, none of the mean RCI or SRB values broke the oft-used threshold of ±1.645 of a clinically meaningful change. One interpretation of these values is that practice effects in patients with MCI are comparable to those in age-matched peers. For example, using Iverson's (2001) practice corrected RCI, 90% of Duff's (2014) sample had improved by 1.7–6.1 points on the Delayed Recall trial of the HVLT–R across one week. The vast majority of the current sample also showed one-week practice effects within that range. However, one of the main advantages of RCIs and SRBs is that they allow us to identify specific individuals who do exceed these thresholds. For example, even though the dependent *t* test indicated that patients with MCI significantly improved by over two words on average on the Delayed Recall trial of the HVLT–R when they were assessed a week later, the RCI would be able to identify the 3.4% (*n* = 2) of the current sample that had a *z* score ≥ 1.645 on this cognitive score (or a raw score improvement of more than 6.1 points). Interestingly, neither the simple nor the complex SRB identified any subjects who significantly improved on this cognitive score. Conversely, if you were interested in finding only those subjects who performed significantly below expectations on practice effects on this cognitive score, the two SRB formulae could identify the 34–36% of subjects that performed ≤ –1.645 on the HVLT–R Delayed Recall. These change formulae make such methods valuable for clinical trials (e.g., using practice effects to enrich samples with patients showing abnormally low scores). Clinically, smaller than expected practice effects may portend a declining trajectory (Duff et

al., 2011), poorer response to an intervention (Duff et al., 2010), or greater risk of Alzheimer's-related pathology (Duff et al., 2014; Galvin et al., 2005; Mormino et al., 2014).

For those less familiar with these various change formulae, a case example of their utility may be helpful. A 79-year-old female with 15 years of education who has been diagnosed with MCI was tested twice across one week, and her raw scores on the HVLT–R Delayed Recall are presented in the upper row of Table 6. It is seen that she improves on this verbal memory measure by 2 raw score points. This amount of improvement is comparable to that observed in Duff (2014), as seen in the Delayed Recall z scores of the RCI of −0.08 (47th percentile), simple SRB of −0.06 (47th percentile), and complex SRB of −0.11 (46th percentile). In general, this patient shows "normal," "typical," "average," or "expected" practice effects across one week, which may suggest a better cognitive future, better response to treatments, and less Alzheimer's pathology. Conversely, if this same patient showed the one-week retest scores as seen in the lower row of Table 6, then different conclusions might be drawn. She shows exactly the same raw scores on Delayed Recall between her baseline and one-week visits. Compared to the sample in Duff (2014), this change is poorer on Delayed Recall (RCI: −0.92, 18th percentile; simple SRB: −1.12, 13th percentile; complex SRB: −1.18, 12th percentile). In general, although none of these values cross the $z \pm 1.645$ threshold, this patient still shows "less normal," "less typical," "low average," or "smaller than expected" practice effects across one week, which may point towards a poorer cognitive future, poorer response to treatments, and greater Alzheimer's pathology.

In comparing the three change formulae, differences seemed to emerge. Across most cognitive variables, the complex SRB yielded the highest mean absolute values, indicating the most change. As seen in Table 4, the complex SRB also seemed to indicate more change in this sample, with 6 of 7 cognitive test scores showing statistically significant one-sample chi-square results of fewer "stable" cases (compared to 3/7 for the simple SRB and 0/7 for the RCI). Similarly, when directly comparing the three change formulae with Cochran's Q (Table 5), the complex SRB tended to show fewer "stable" cases than either the simple SRB or the RCI. Although there remains no consensus as to which change formulae are preferred to detect reliable and clinically meaningful change (Frerichs & Tuokko, 2005; Heaton et al., 2001; Hinton-Bayre, 2010; Maassen et al., 2009; Temkin et al., 1999), the current results lend some preliminary support in a clinical sample that the complex SRB identifies more change than simpler methods. To our knowledge, this is the first study to examine multiple change methods in patients with amnestic MCI, which may explain the differences with the existing literature. But it would be an inaccurate conclusion to say that the complex SRB is necessarily the "best" formula. Without some external criterion (e.g., receiving a diagnosis of dementia, hippocampal atrophy, transition to an assisted living facility), we can only say that complex SRB yields higher rates of change. Future research is clearly needed to understand which change formulae are most strongly linked to real-world outcomes. For example, Frerichs and Tuokko (2005) found that multiple change formulae were related to conversion from intact to dementia in a large sample of older adults.

In addition to needing some external criterion with which to compare the different change formulae, there are some other limitations of the current study. First, these results only

inform us about change in patients with amnestic MCI, as we did not have other groups (e.g., cognitively intact, Alzheimer's disease). Prior studies have examined these other phases along the continuum of late-life cognition (Bläsi et al., 2009; Duff et al., 2012; Gavett, Ashendorf, & Gurnani, 2015), and the current results would not necessarily generalize to those other groups. Our comparison sample from Duff (2014) was composed of subjects classified as cognitively intact and amnestic MCI. In some ways, it may be atypical to compare the current subjects to a cognitively heterogeneous group, as it is more typical to compare a clinical cohort to a neurologically healthy comparison group (Temkin et al., 1999). However, there may be advantages of using a heterogeneous comparison sample (e.g., wider range of baseline and follow-up scores that may increase generalizability of findings), as noted by Heaton et al. (2001). Second and similarly, these results would not likely generalize to differences in the experimental design of this study. For example, applying these findings to different memory tests (e.g., California Verbal Learning Test, Rey Osterrieth Complex Figure Test) or different retest intervals (e.g., 6, 12, or 24 months) is not advisable, as Calamia et al. (2012) has demonstrated that practice effects are influenced by such factors. Although the current sample was evenly split between males and females, the sample in Duff (2014) was not, being over 80% female. It is possible that gender differences between the samples may have affected change scores across one week, and gender should be considered in future studies on change. Third, with only 58 subjects in our current sample, our power may have been relatively low to detect subtler practice effects in the processing-speed tests. However, practice effects are not universally found on such cognitive measures (Basso, Bornstein, & Lang, 1999). Fourth, ceiling effects need to be considered when assessing change, especially when test scores have limited ranges. For example, on HVLT–R Delayed Recall, it would be exceedingly difficult for an individual who obtained a score of 7 at baseline (the mean score in Duff's sample) to show enough improvement to surpass the +1.645 point that typically indicates "improvement above and beyond practice effects," even if he or she got all 12 words correct after one week (e.g., RCI plus practice effects $z = 1.17$, simple SRB $z = 1.52$, complex SRB $z = 1.65$ if the individual was 87 years old). Floor effects, however, appear less problematic. If an individual obtained a score of 0 at baseline on the HVLT–R Delayed Recall, a score of 0 after one week would still be suggestive of less than expected change in this individual on some change metrics (e.g., RCI plus practice effects $z = -0.92$, simple SRB $z = -2.81$, complex SRB $z = -3.12$ if the individual was 75 years old). Despite these limitations, the current results show that individuals with MCI do demonstrate practice effects across one week, although the amount of improvement tends to be smaller than expected relative to their peers. These results also validate the RCI and SRB change norms developed by Duff (2014), which can be used in similar patients to more accurately quantify practice effects across brief intervals. Finally, these findings highlight that practice effects remain a complex construct, worthy of continued investigation in diverse clinical conditions.

Acknowledgments

solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health.

References

Basso MR, Bornstein RA, Lang JM. Practice effects on commonly used measures of executive function across twelve months. The Clinical Neuropsychologist. 1999; 13:283–292. DOI: 10.1076/clin.13.3.283.1743 [PubMed: 10726600]

Benedict, RHB. Brief visuospatial memory test —Revised: Professional manual. Lutz, FL: Psychological Assessment Resources; 1997.

Bläsi S, Zehnder AE, Berres M, Taylor KI, Spiegel R, Monsch AU. Norms for change in episodic memory as a prerequisite for the diagnosis of mild cognitive impairment (MCI). Neuropsychology. 2009; 23:189–200. DOI: 10.1037/a0014079 [PubMed: 19254092]

Brandt, J., Benedict, RHB. Hopkins verbal learning test—Revised: Professional manual. Lutz, FL: Psychological Assessment Resources; 2001.

Britt WG III, Hansen AM, Bhaskerrao S, Larsen JP, Petersen F, Dickson A, … Kirsch WM. Mild cognitive impairment: Prodromal Alzheimer's disease or something else? Journal of Alzheimer's Disease. 2011; 27:543–551. DOI: 10.3233/JAD-2011-110740

Calamia M, Markon K, Tranel D. Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. The Clinical Neuropsychologist. 2012; 26:543–570. DOI: 10.1080/13854046.2012.680913 [PubMed: 22540222]

Chelune GJ, Naugle RI, Lüders H, Sedlak J, Awad IA. Individual change after epilepsy surgery: Practice effects and base-rate information. Neuropsychology. 1993; 7:41–52. DOI: 10.1037/0894-4105.7.1.41

Chiu EC, Koh CL, Tsai CY, Lu WS, Sheu CF, Hsueh IP, Hsieh CL. Practice effects and test-re-test reliability of the five digit test in patients with stroke over four serial assessments. Brain Injury. 2014; 28:1726–1733. DOI: 10.3109/02699052.2014.947618 [PubMed: 25188016]

Cooper DB, Epker M, Lacritz L, Weiner M, Rosenberg RN, Honig L, Cullum CM. Effects of practice on category fluency in Alzheimer's disease. The Clinical Neuropsychologist. 2001; 15:125–128. DOI: 10.1076/clin.15.1.125.1914 [PubMed: 11778573]

Cooper DB, Lacritz LH, Weiner MF, Rosenberg RN, Cullum CM. Category fluency in mild cognitive impairment: Reduced effect of practice in test-retest conditions. Alzheimer Disease & Associated Disorders. 2004; 18:120–122. DOI: 10.1097/01.wad.0000127442.15689.92 [PubMed: 15494616]

Darby D, Maruff P, Collie A, McStephen M. Mild cognitive impairment can be detected by multiple assessments in a single day. Neurology. 2002; 59:1042–1046. DOI: 10.1212/WNL.59.7.1042 [PubMed: 12370459]

Duff K. Evidence-based indicators of neuropsy-chological change in the individual patient: Relevant concepts and methods. Archives of Clinical Neuropsychology. 2012; 27:248–261. DOI: 10.1093/arclin/acr120 [PubMed: 22382384]

Duff K. One-week practice effects in older adults: Tools for assessing cognitive change. The Clinical Neuropsychologist. 2014; 28:714–725. DOI: 10.1080/13854046.2014.920923 [PubMed: 24882553]

Duff K, Beglinger LJ, Moser DJ, Schultz SK, Paulsen JS. Practice effects and outcome of cognitive training: Preliminary evidence from a memory training course. The American Journal of Geriatric Psychiatry. 2010; 18:91.doi: 10.1097/JGP.0b013e3181b7ef58 [PubMed: 20104658]

Duff K, Beglinger LJ, Schultz SK, Moser DJ, McCaffrey RJ, Haase RF, … Paulsen J. Huntington's Study Group. Practice effects in the prediction of long-term cognitive outcome in three patient samples: A novel prognostic index. Archives of Clinical Neuropsychology. 2007; 22:15–24. DOI: 10.1016/j.acn.2006.08.013 [PubMed: 17142007]

Duff K, Callister C, Dennett K, Tometich D. Practice effects: A unique cognitive variable. The Clinical Neuropsychologist. 2012; 26:1117–1127. DOI: 10.1080/13854046.2012.722685 [PubMed: 23020261]

Duff K, Foster NL, Hoffman JM. Practice effects and amyloid deposition: Preliminary data on a method for enriching samples in clinical trials. Alzheimer Disease & Associated Disorders. 2014; 28:247–252. DOI: 10.1097/WAD.0000000000000021 [PubMed: 24614265]

Duff K, Lyketsos CG, Beglinger LJ, Chelune G, Moser DJ, Arndt S, … McCaffrey RJ. Practice effects predict cognitive outcome in amnestic mild cognitive impairment. The American Journal of Geriatric Psychiatry. 2011; 19:932–939. DOI: 10.1097/JGP.0b013e318209dd3a [PubMed: 22024617]

Frerichs RJ, Tuokko HA. A comparison of methods for measuring cognitive change in older adults. Archives of Clinical Neuropsychology. 2005; 20:321–333. DOI: 10.1016/j.acn.2004.08.002 [PubMed: 15797168]

Galvin JE, Powlishta KK, Wilkins K, McKeel DW, Xiong C, Grant E, … Morris JC. Predictors of preclinical Alzheimer disease and dementia: A clinicopathologic study. Archives of Neurology. 2005; 62:758–765. DOI: 10.1001/archneur.62.5.758 [PubMed: 15883263]

Gavett BE, Ashendorf L, Gurnani AS. Reliable change on neuropsychological tests in the uniform data set. Journal of the International Neuropsychological Society. 2015; 21:558–567. DOI: 10.1017/S1355617715000582 [PubMed: 26234918]

Hassenstab J, Ruvolo D, Jasielec M, Xiong C, Grant E, Morris JC. Absence of practice effects in preclinical Alzheimer's disease. Neuropsychology. 2015; 29:940–948. DOI: 10.1037/neu0000208 [PubMed: 26011114]

Heaton RK, Temkin N, Dikmen S, Avitable N, Taylor MJ, Marcotte TD, Grant I. Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. Archives of Clinical Neuropsychology. 2001; 16:75–91. DOI: 10.1093/arclin/16.1.75 [PubMed: 14590193]

Helkala EL, Kivipelto M, Hallikainen M, Alhainen K, Heinonen H, Tuomilehto J, … Nissinen A. Usefulness of repeated presentation of mini-mental state examination as a diagnostic procedure –A population-based study. Acta Neurologica Scandinavica. 2002; 106:341–346. DOI: 10.1034/j.1600-0404.2002.01315.x [PubMed: 12460138]

Hinton-Bayre AD. Deriving reliable change statistics from test–retest normative data: Comparison of models and mathematical expressions. Archives of Clinical Neuropsychology. 2010; 25:244–256. DOI: 10.1093/arclin/acq008 [PubMed: 20197293]

Iverson GL. Interpreting change on the WAIS-III/WMS-III in clinical samples. Archives of Clinical Neuropsychology. 2001; 16:183–191. DOI: 10.1093/arclin/16.2.183 [PubMed: 14590186]

Linebach, JA., Tesch, BP., Kovacsiss, LP. Nonparametric statistics for applied research. New York, NY: Springer; 2014.

Maassen GH, Bossema E, Brand N. Reliable change and practice effects: Outcomes of various indices compared. Journal of Clinical and Experimental Neuropsychology. 2009; 31:339–352. DOI: 10.1080/13803390802169059 [PubMed: 18618359]

Machulda MM, Pankratz VS, Christianson TJ, Ivnik RJ, Mielke MM, Roberts RO, … Petersen RC. Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. The Clinical Neuropsychologist. 2013; 27:1247–1264. DOI: 10.1080/13854046.2013.836567 [PubMed: 24041121]

Mathews M, Abner E, Kryscio R, Jicha G, Cooper G, Smith C, … Schmitt FA. Diagnostic accuracy and practice effects in the national Alzheimer's coordinating center uniform data set neuropsychological battery. Alzheimer's & Dementia. 2014; 10:675–683. DOI: 10.1016/j.jalz.2013.11.007

McCaffrey, RJ.Duff, K., Westervelt, HJ., editors. Practitioner's guide to evaluating change with neuropsychological assessment instruments. New York, NY: Springer Science & Business Media; 2000.

McSweeny AJ, Naugle RI, Chelune GJ, Lüders H. "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. The Clinical Neuropsychologist. 1993; 7:300–312. DOI: 10.1080/13854049308401901

Mormino EC, Betensky RA, Hedden T, Schultz AP, Amariglio RE, Rentz DM, … Sperling RA. Synergistic effect of β-amyloid and neurode-generation on cognitive decline in clinically normal individuals. JAMA Neurology. 2014; 71:1379–1385. DOI: 10.1001/jamaneurol.2014.2031 [PubMed: 25222039]

Reitan, RM. Trail making test: Manual for administration and scoring. Tucson, AZ: Reitan Neuropsychology Laboratory; 1992.

Rogers JM, Fox AM, Donnelly J. Impaired practice effects following mild traumatic brain injury: An event-related potential investigation. Brain Injury. 2015; 29:343–351. DOI: 10.3109/02699052.2014.976273 [PubMed: 25436891]

Rosenthal, R. Meta-analytic procedures for social research. Newbury Park, CA: SAGE Publications; 1991.

Schrijnemaekers AMC, de Jager CA, Hogervorst E, Budge MM. Cases with mild cognitive impairment and Alzheimer's disease fail to benefit from repeated exposure to episodic memory tests as compared with controls. Journal of Clinical and Experimental Neuropsychology. 2006; 28:438–455. DOI: 10.1080/13803390590935462 [PubMed: 16618630]

Smith, A. Symbol digit modalities test [manual]. Torrance, CA: Western Psychological Services; 1973.

Suchy Y, Kraybill ML, Franchow E. Practice effect and beyond: Reaction to novelty as an independent predictor of cognitive decline among older adults. Journal of the International Neuropsychological Society. 2011; 17:101–111. DOI: 10.1017/S135561771000130X [PubMed: 21073771]

Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: A comparison of four models. Journal of the International Neuropsychological Society. 1999; 5:357–369. DOI: 10.1017/S1355617799544068 [PubMed: 10349298]

Wilkinson, GS., Robertson, GJ. WRAT 4: Wide range achievement test; professional manual. Lutz, FL: Psychological Assessment Resources; 2006.

Yan JH, Dick MB. Practice effects on motor control in healthy seniors and patients with mild cognitive impairment and Alzheimer's disease. Aging, Neuropsychology, and Cognition. 2006; 13:385–410. DOI: 10.1080/138255890969609

Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, Leirer VO. Development and validation of a geriatric depression screening scale: A preliminary report. Journal of Psychiatric Research. 1983; 17:37–49. DOI: 10.1016/0022-3956(82)90033-4

Zehnder AE, Bläsi S, Berres M, Spiegel R, Monsch AU. Lack of practice effects on neuropsychological tests as early cognitive markers of Alzheimer disease? American Journal of Alzheimer's Disease and Other Dementias. 2007; 22:416–426. DOI: 10.1177/1533317507302448

Table 1

Reliable change scores and their formulae.

Reliable change score	Formula
RCI	$z_{RCI} = (T_2 - T_1) - (M_2 - M_1)/SED$ $SED = [(S_1 \ 1 - r_{12})^2 + (S_2 \ 1 - r_{12})^2]$
"simple" SRB	$T_2' = c + bT_1$ $z_{SRB} = T_2 - T_2'/SEE$
"complex" SRB	$T_2' = c + bT_1$ + other variables in the model $z_{SRB} = T_2 - T_2'/SEE$

Note. RCI = reliable change index; SRB = standardized regression-based model; T_1 = score at Time 1; T_2 = score at Time 2; S_1 = standard deviation at Time 1; S_2 = standard deviation at Time 2; r_{12} = correlation between Time 1 and 2 scores; M_1 = control group mean at Time 1; M_2 = control group mean at Time 2; SED = standard error of the difference; b = slope of the regression model (beta coefficient for T_1); c = intercept of the regression model (constant); SEE = standard error of the estimate of the regression model; T_2' = predicted score at Time 2 based on regression model.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Cognitive scores at baseline and one-week visits and raw practice effects.

Cognitive scores	Baseline	One week	Practice effects	r	d
HVLT–R					
Total recall	17.5 (5.5)	20.8 (6.8)	3.3 (4.0) *	.80	0.81
Delayed recall	3.5 (3.3)	5.5 (4.1)	2.1 (2.1) *	.87	1.01
BVMT–R					
Total recall	10.3 (5.7)	17.1 (9.0)	6.8 (5.4) *	.82	1.26
Delayed recall	3.8 (3.0)	6.0 (3.6)	2.2 (1.8) *	.87	1.22
TMT–A	43.7 (16.8)	43.2 (17.4)	0.5 (11.7)	.77	−0.04
TMT–B	145.0 (85.1)	137.1 (83.6)	10.2 (46.0)	.85	0.22
SDMT	34.9 (10.5)	35.0 (12.2)	0.1 (5.7)	.88	0.01

Note. Practice effects are one-week raw score – baseline raw score; r = correlation between baseline and one-week scores; d = Cohen's d of practice effects; HVLT–R = Hopkins Verbal Learning Test–Revised; BVMT–R = Brief Visuospatial Memory Test–Revised; TMT = Trail Making Test; SDMT = Symbol Digit Modalities Test.

*
$p < .05$.

Table 3

RCI and SRB values for sample.

Cognitive scores	RCI	Simple SRB	Complex SRB	Post hoc
HVLT–R				
Total recall	−0.28 (1.0)	−0.60 (1.1)	−0.60 (1.1)	1 < 2,3
Delayed recall	−0.05 (0.9)	−0.89 (1.4)	−1.09 (1.5)	1 < 2 < 3
BVMT–R				
Total recall	−0.26 (1.0)	−0.23 (1.0)	−0.56 (1.1)	2 < 1 < 3
Delayed recall	−0.05 (0.8)	−0.29 (1.0)	−0.26 (1.2)	1 < 2,3
TMT–A	−0.33 (1.1)	−0.35 (1.1)	−0.49 (1.1)	1,2 < 3
TMT–B	−0.06 (1.2)	−0.35 (1.3)	−0.58 (1.4)	1 < 2 < 3
SDMT	−0.43 (1.1)	−0.44 (1.1)	−0.74 (1.2)	1,2 < 3

Note. HVLT–R = Hopkins Verbal Learning Test–Revised; BVMT–R = Brief Visuospatial Memory Test–Revised; TMT = Trail Making Test; SDMT = Symbol Digit Modalities Test; RCI = reliable change index; SRB = standardized regression based formula; post hoc = post hoc comparison between groups (1 = RCI, 2 = simple SRB, 3 = complex SRB) following repeated measures analysis of variance (ANOVA). The signs for TMT–A and TMT–B were reversed so that values for all measures went in the same direction (i.e., negative values indicate smaller than expected practice effects).

Table 4

Percentage of cases that decline, remain stable, or improve.

Cognitive scores	RCI			Simple SRB			Complex SRB		
	Decline	Stable	Improve	Decline	Stable	Improve	Decline	Stable	Improve
HVLT–R									
Total recall	6.9	89.7 $p = .93$	3.4	19.0	79.3 $p = .007$	1.7	19.0	79.3 $p = .007$	1.7
Delayed recall	0	96.6 $p = .10$	3.4	34.5	65.5 $p < .001$	0	36.2	63.8 $p < .001$	0
BVMT–R									
Total recall	6.9	89.7 $p = .93$	3.4	5.2	91.4 $p = .73$	3.4	17.2	77.6 $p = .002$	5.2
Delayed recall	1.7	94.8 $p = .22$	3.4	12.1	86.2 $p = .34$	1.7	15.5	79.3 $p = .007$	5.2
TMT–A	3.4	86.2 $p = .34$	10.3	3.4	84.5 $p = .16$	12.1	13.8	82.8 $p = .07$	3.4
TMT–B	8.9	83.9 $p = .13$	7.1	3.6	76.8 $p = .004$	19.6	17.9	78.6 $p = .001$	3.6
SDMT	10.3	87.9 $p = .60$	1.7	10.3	86.2 $p = .16$	3.4	19.0	75.9 $p < .001$	5.2

Note. See text for definition of decline, stable, and improve. RCI = reliable change index; SRB = standardized regression-based formula; HVLT–R = Hopkins Verbal Learning Test–Revised; BVMT–R = Brief Visuospatial Memory Test–Revised; TMT = Trail Making Test; SDMT = Symbol Digit Modalities Test; *p*-values are reported for one-sample chi-square of the percentage of stable cases that differ from the expected value of 90%.

Table 5

Percentage of cases that were stable or not stable.

Cognitive scores	RCI		Simple SRB		Complex SRB		
	Stable	Not stable	Stable	Not stable	Stable	Not stable	Post hoc
HVLT–R							
Total recall	10.3	89.7	20.7	79.3	20.7	79.3	nd
Delayed recall	3.4	96.6	34.5	65.5	36.2	63.8	1 > 2 > 3
BVMT–R							
Total recall	10.3	89.7	8.6	91.4	22.4	77.6	2 > 3
Delayed recall	5.2	94.8	13.8	86.2	20.7	79.3	1 > 3
TMT–A	13.8	86.2	15.5	84.5	17.2	82.8	nd
TMT–B	16.1	83.9	23.2	76.8	21.4	78.6	nd
SDMT	12.1	87.9	13.8	86.2	24.1	75.9	1,2 > 3

Note. See text for definition of stable and not stable. RCI = reliable change index; SRB = standardized regression-based formula; HVLT–R = Hopkins Verbal Learning Test–Revised; BVMT–R = Brief Visuospatial Memory Test–Revised; TMT = Trail Making Test; SDMT = Symbol Digit Modalities Test. Post hoc = McNemar post hoc comparison between formulae (1 = RCI, 2 = simple SRB, 3 = complex SRB) showing which formula had the most stable cases. nd = no difference between formulae.

Table 6

Case example showing expected practice effects on cognitive scores on the HVLT–R Delayed Recall.

Baseline	One week	RCI	Simple SRB	Complex SRB
7	9	−0.08	−0.06	−0.11
7	7	−0.92	−1.12	−1.18

Note. RCI = reliable change index; SRB = standardized regression based formula; HVLT–R = Hopkins Verbal Learning Test–Revised. Baseline and one-week scores are raw scores. RCI, simple SRB, and complex SRB scores are *z* scores, where negative *z* scores indicate smaller than expected change relative to Duff (2014).